Can Evolutionary Rate Matrices be Estimated from Allele Frequencies?

Conrad J. Burden
Mathematical Sciences Institute, Australian National University, Canberra, ACT 2601, Australia

Keywords: Evolutionary Rate, Wright-Fisher Model, Fokker-Plank Equation.

Abstract: This paper is a work in progress in which aims to combine the principles of population genetics and continuous-time Markovian evolutionary models to estimate evolutionary rate matrices from the current observed state of a single genome. A model is proposed in which sections of the genome which are not susceptible to natural selection are considered to be a statistical ensemble of individual genomic sites. Each site is a representative from a stationary distribution of allele frequencies $0 \leq \theta \leq 1$ within the population. Simulations of this distribution via a finite-state Markov model based on a finite effective population size are compared with the stationary solution to the continuum Fokker-Planck equation. Parameters of the evolutionary rate matrix introduced via mutation rates within the Fokker-Planck equation are estimated for simulated data in a number of exploratory examples.

1 INTRODUCTION

The rapidly reducing cost of high-throughput sequencing now allows for the acquisition of genome-wide data profiling allele frequencies within populations across large numbers of polymorphic sites (Lynch, 2009). This paper is an exploratory analysis of the feasibility of estimating evolutionary rate matrices solely from the current observed state of allele frequencies within a genome.

The evolutionary rate matrix at a given genomic site is known to depend strongly on the site's context, that is, the nucleotide content of its flanking bases (Nevarez et al., 2010; Zhao and Boerwinkle, 2002). Herein we will assume the available data to be sufficiently extensive that (1) individual mutation rates can be fitted independently for each context, and (2) that we can restrict ourselves to sites expected to evolve via a "neutral" theory in which the effects of natural selection can be ignored. Within these constraints, the assumed data consists of a set of 4 numbers $(\theta_A, \theta_C, \theta_G, \theta_G)$ at any given site in the genome giving the relative abundance $0 \leq \theta_a \leq 1$ of nucleotide $a \in \{A, C, G, T\}$ within the population at that site. For the vast majority of sites this vector is observed to be (immeasurably close to) one of $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$ or $(0,0,0,1)$. Sites for which two or more components are non-zero are referred to as single nucleotide polymorphisms (SNPs), and for most SNPs only two components are observed to be non-zero. The nucleotides for which θ_a is non-zero at a given SNP are referrred to as alleles.

A similar approach to that set out here has also been taken by Messer (Messer, 2009). However, Messer's approach differs in that he restricts the data to alleles occurring with low frequency in the population, that is θ close to 0 or 1, in order to reduce the effects of selection. In our approach we assume the set of genomic sites can be reduced to those not subject to selection and use the entire range of allele frequencies.

2 THE MODEL

Our starting point is a discrete-time Markov model which combines two fundamental ideas of population genetics.

The first of these is the discrete stochastic model of genetic drift (see for instance (Ewens, 2004), Chapter 3), defined by a square, time-independent transition matrix p_{ij} defined as follows: If, at a SNP within the genome with two alleles A_1 and A_2, $Y(\tau)$ is the number of individuals in a diploid population of size N with the allele A_1 at time-step (or generation) $\tau = 1, 2, \ldots$, then

$$\text{Prob}\,(Y(\tau+1) = j | Y(\tau) = i) = p_{ij}, \quad i, j = 0, 1, \ldots, M,$$

where $M = 2N$. The canonical model generally considered is the Wright-Fisher model (Wright, 1931) for

J. Burden C..
Can Evolutionary Rate Matrices be Estimated from Allele Frequencies?.
DOI: 10.5220/0005253701830188
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2015), pages 183-188
ISBN: 978-989-758-070-3

a monoecious population, for which

$$P_{ij} = \binom{M}{j} \left(\frac{i}{M}\right)^j \left(1-\frac{i}{M}\right)^{(M-j)}.$$

This view of genetic drift readily generalises to an alphabet $\mathcal{L} = \{A,C,G,T\}$ and a vector of random variables $\mathbf{Y}(\tau) = (Y_A, Y_C, Y_G, Y_T)$ with a probability distribution

$$\text{Prob}\,(\mathbf{Y}(\tau) = \mathbf{i}) = \phi(i_A, i_C, i_G, i_T; \tau),$$

where

$$\sum_{a\in\mathcal{L}} i_a = M, \quad \text{and } i_a = 0,\ldots,M. \tag{1}$$

In general, for an alphabet of size d, the vector ϕ has $\binom{M+d-1}{d-1}$ components. We interpret this distribution as the relative frequency of genomic sites at which alleles are present at the population frequency $\mathbf{i}/M$. Most of the components of ϕ will be very close to zero in practice as SNPs with more than two alleles are extremely rare within the genome. Furthermore since the vast majority of genomic sites are not SNPs, the distribution will be heavily dominated by the components $\phi(M,0,0,0)$, $\phi(0,M,0,0)$, etc. The Wright-Fisher model thus generalises to

$$P_{\mathbf{ij}} = \text{Prob}\,(\mathbf{Y}(\tau+1) = \mathbf{j}|\mathbf{Y}(\tau) = \mathbf{i}) = \binom{M}{j_A j_C j_G j_T} \left(\frac{i_A}{M}\right)^{j_A} \left(\frac{i_C}{M}\right)^{j_C} \left(\frac{i_G}{M}\right)^{j_G} \left(\frac{i_T}{M}\right)^{j_T}.$$

The second fundamental idea is that genomic mutations are modelled via a site-independent Markov transition matrix

$$\Pi(t) = exp(tQ)$$

where the elements q_{ab} of Q, satisfying $\sum_{b\in\mathcal{L}} q_{ab} = 0$, represent the instantaneous mutation rate from allele a to allele b. Our ultimate aim is to estimate Q from allele frequencies θ within the population at each of the sites within the restricted set of genomic sites described in the Introduction.

To make contact with the above model of genetic drift, it is instructive to re-visit the Wright-Fisher model. Assuming a two-step process in which inheritance is followed by mutation at each generation, the transition matrix of the Wright-Fisher model becomes ((Ewens, 2004), Chapter 3)

$$P_{ij} = \binom{M}{j} \psi(i)^j (1-\psi(i))^{M-j}, \tag{2}$$

where

$$\psi(i) = \frac{i}{M}(1-u) + \left(1-\frac{i}{M}\right)v,$$

where u is the probability of mutation from allele A_1 to allele A_2 and v the probability of mutation from A_2 to A_1 in one generation. Here u and v are assumed to be $O(1/M)$. If we scale the continuous time according to $t = \tau/M$, this corresponds to a mutation Markov matrix over one generation of

$$\Pi(1/M) = \begin{pmatrix} 1-u & u \\ v & 1-v \end{pmatrix} + O(M^{-2}),$$

where the first row and column correspond to allele A_1 and the second row and column correspond to allele A_2.

By making the analogous substitutions in the generalised Wright-Fisher model, and defining u_{ab}, $a,b \in \mathcal{L}$, to be the rate of mutation from allele a to allele b in one generation, one arrives at

$$P_{\mathbf{ij}} = \text{Prob}\,(\mathbf{Y}(\tau+1) = \mathbf{j}|\mathbf{Y}(\tau) = \mathbf{i}) = \binom{M}{j_A j_C j_G j_T} \prod_{a\in\{A,C,G,T\}} \psi(\mathbf{i},a)^{j_a}, \tag{3}$$

where

$$\psi(\mathbf{i},a) = \frac{i_a}{M}\left(1-\sum_{b\neq a} u_{ab}\right) + \sum_{b\neq a} \frac{i_b}{M} u_{ba}.$$

The off-diagonal elements of the instantaneous evolutionary rate matrix are

$$q_{ab} = M u_{ab}.$$

The observed data, as described in the Introduction are assumed to correspond to the stationary distribution of the matrix $P_{\mathbf{ij}}$, thus allowing for an estimate of the parameters q_{ab}.

3 TOY MODEL: 2-LETTER ALPHABET

To explore the feasibility of estimating parameters of the rate matrix from a data set, we begin with the case of a 2-letter alphabet, described by the evolution matrix P_{ij} defined by Equation (2). In this case the only parameters of the model are the off-diagonal elements of the continuous-time evolutionary rate matrix, $q_{12} = Mu$ and $q_{21} = Mv$ and (twice the) effective population M. Figure 1 shows the stationary distribution

$$\phi(i) = \text{Prob}\,(Y(\tau=\infty) = i), \tag{4}$$

obtained numerically for the parameter values stated in the figure caption. The parameters q_{12} and q_{21} have been chosen to some extent so the the distribution is dominated by the end points $i = 0$ and M to mimic the behaviour of real genomes in which the vast majority

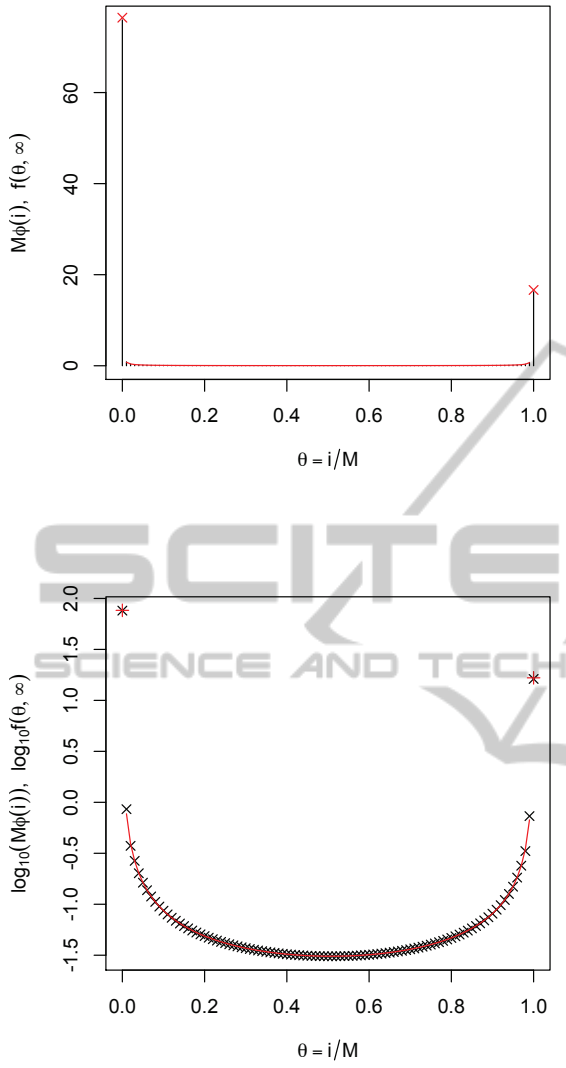

Figure 1: Numerical stationary distribution $\phi(i)$ of the Wright-Fisher model with two-way mutations for the case $M = 100$, $q_{12} = 0.02$ and $q_{21} = 0.005$ plotted on a linear (top) and logarithmic (bottom) scale. Also shown in red is the stationary solution $f(\theta,\infty)$ of the continuum $M \to \infty$ Fokker-Planck equation with the end points of the distribution approximated as explained in the text. The extra factor of M in the vertical axis scale set to enable a comparison with the continuum stationary distribution.

of sites are not SNPs. This point is discussed further below.

The effective population size of 100 in this numerical simulation is of course unrealistically low, and we next consider the limit $M \to \infty$. There is a well established literature on building partial differential equations, known as Fokker-Planck or forward Kolmogorov diffusion equations, to describe the time evolution of the probability density of allele frequencies θ for a given SNP. The continuum Fokker-Planck equation for the probability density $f(\theta,t)$ is obtained by setting $\theta = i/M$, $t = \tau/M$ and taking the limit $M \to \infty$ of the discrete model to obtain (Ewens, 2004)

$$\frac{\partial f}{\partial t} = -\frac{\partial}{\partial \theta}(a(\theta) f(\theta,t)) + \frac{1}{2}\frac{\partial^2}{\partial \theta^2}(b(\theta) f(\theta,t)), \quad (5)$$

where, for the current model,

$$a(\theta) = -q_{12}\theta + q_{21}(1-\theta),$$
$$b(\theta) = \theta(1-\theta).$$

Setting the time derivative to zero to obtain the stationary distribution and normalising so that $\int_0^1 f(\theta)d\theta = 1$ yields the well-known beta distribution

$$f(\theta,\infty) = B(2q_{12}, 2q_{21})\theta^{2q_{21}-1}(1-\theta)^{2q_{12}-1}. \quad (6)$$

where

$$B(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

This function is superimposed over the finite M solution over the range $1/M \le \theta \le 1 - 1/M$ in Figure 1, and clearly illustrates that the continuum limit is approached very rapidly. To obtain the end points plotted at $\theta = 0$ and 1 we used the approximations

$$\begin{aligned} \text{Prob}\,(Y(\infty) = 0) &\approx \int_0^{1/M} f(\theta,\infty)d\theta \\ &\approx \frac{B(2q_{12}, 2q_{21})}{2q_{21}M^{2q_{21}-1}}, \end{aligned} \quad (7)$$

and

$$\begin{aligned} \text{Prob}\,(Y(\infty) = M) &\approx \int_{1-1/M}^{1} f(\theta,\infty)d\theta \\ &\approx \frac{B(2q_{12}, 2q_{21})}{2q_{12}M^{2q_{12}-1}}. \end{aligned} \quad (8)$$

In summary, we see that the continuum limit $M \to \infty$ is reached rapidly, the precise value of M is in some sense irrelevant, and that for practical purposes its role is to provide a lower bound $1/M$ on the frequency of a rare allele before a genomic site is deemed to be a SNP.

4 PARAMETER ESTIMATION

We consider next the problem of estimating the parameters of the rate matrix for the toy model described in the previous section.

4.1 Entire Population Surveyed

We start with the following artificially contrived situation concerning a hypothetical life form whose

Table 1: Mutation rates $\hat{q}_{12}$ and $\hat{q}_{21}$ estimated from synthetic data generated from the two-letter alphabet model assuming the entire population of M chromosomes is genotyped at n_{site} genomic sites.

M	q_{12}	q_{21}	n_{site}	$\hat{q}_{12}$	$\hat{q}_{21}$
100	0.02	0.005	10^3	0.0194	0.00462
			10^4	0.0207	0.00519
			10^5	0.0206	0.00504
100	0.05	0.005	10^3	0.0508	0.00485
			10^4	0.0541	0.00488
			10^5	0.0529	0.00509
200	0.02	0.005	10^3	0.0253	0.00607
			10^4	0.0209	0.00517
			10^5	0.0206	0.00509

genome is built from a two-letter alphabet: Suppose we have a small monoecious, diploid population of effective size $M/2$, and we are able to genotype the entire population of M chromosomes at a complete set of n_{site} independent genomic sites, each assumed chosen to be a site not susceptible to selective pressures. The data from which we are to estimate the rate parameters q_{12} and q_{21} consist of a set of observed allele frequencies $\theta_1, \theta_2, \ldots, \theta_{n_{\text{site}}}$ each taking a value in the set $\{0, 1/M, 2/M, \ldots, 1\}$. For the vast majority of these sites we will observe $\theta = 0$ or 1, however it is important to retain these non-SNP sites in the data.

Table 1 gives maximum likelihood estimates $\hat{q}_{12}$ and $\hat{q}_{21}$ of mutation rates from synthetic data generated from the numerically determined stationary distribution of the Wright-fisher model with mutation, namely Equation (4). The log likelihood is calculated from these data using the continuum limit stationary distribution, Equations (6) to (8).

4.2 Population Sampled

More realistically one expects the effective population to be large, and that the data will consist of a relatively small read coverage at each site obtained by sequencing a sample of the population. We will assume a uniform read coverage n_{read} across n_{site} independent genomic sites, each assumed chosen to be a site not susceptible to selective pressures.

In this case the data will consist of a set of numbers $K_1, K_2, \ldots K_{n_{\text{site}}}$ of type A_1 alleles, each taking a value in the set $\{0, 1, \ldots, n_{\text{read}}\}$, observed at the n_{site} genomic sites. At any given site, conditional on the population fraction θ of A_1-type alleles at that site, the observed number of A_1 alleles K will be a binomial random variable:

$$K|\theta \sim \text{bin}(n_{\text{read}}, \theta),$$

where θ has the beta distribution Equation (6). Thus

Table 2: Mutation rates $\hat{q}_{12}$ and $\hat{q}_{21}$ estimated from synthetic data generated from the two-letter alphabet model with rate matrix parameters $q_{12} = 0.02$, $q_{21} = 0.005$ assuming a sample of size n_{read} is genotyped at n_{site} genomic sites.

n_{read}	n_{site}	$\hat{q}_{12}$	$\hat{q}_{21}$
20	10^3	0.0203	0.00579
	10^4	0.0188	0.00474
	10^5	0.0191	0.00484
50	10^3	0.0204	0.00510
	10^4	0.0200	0.00480
	10^5	0.0200	0.00508

Table 3: Same as Table 2, except using rate matrix parameters $q_{12} = 0.002$, $q_{21} = 0.0005$.

n_{read}	n_{site}	$\hat{q}_{12}$	$\hat{q}_{21}$
20	10^3	0.00218	0.000527
	10^4	0.00156	0.000406
	10^5	0.00187	0.000458
50	10^3	0.00138	0.000352
	10^4	0.00160	0.000409
	10^5	0.00192	0.000474

the unconditional distribution of K is beta-binomial (see (Johnson et al., 1992), Chapter 6),

$$\text{Prob}\,(K = k) = \binom{n_{\text{site}}}{k} \frac{B(2q_{12} + k_1, 2q_{21} + n_{\text{site}} - k)}{B(2q_{12}, 2q_{21})}.$$

Tables 2 and 3 give maximum likelihood estimates of $\hat{q}_{12}$ and $\hat{q}_{21}$ for synthetic data generated from the above beta-binomial distribution for realistic read coverages $n_{\text{read}} = 20$ and 40, assuming the number of independent genomic sites sampled is $n_{\text{site}} = 10^3$, 10^4 and 10^5.

In both examples above, we see that reasonable estimates of mutation rates are obtained with experimentally feasible values for n_{site} and n_{read}, and that the estimate generally improves slightly with the the number of genomic sites observed.

5 FULL MODEL WITH HASEGAWA-KISHINO-YANO RATE MATRIX

Finally we demonstrate the form of the stationary solution for the full model with a 4-letter alphabet for the case of the Hasegawa-Kishino-Yano (HKY) rate matrix (Hasegawa et al., 1985):

$$Q = \alpha \begin{pmatrix} \ldots & \beta\pi_C & \pi_G & \beta\pi_T \\ \beta\pi_A & \ldots & \beta\pi_G & \pi_T \\ \pi_A & \beta\pi_C & \ldots & \beta\pi_T \\ \beta\pi_A & \pi_C & \beta\pi_G & \ldots \end{pmatrix},$$

with the diagonal elements set so that each row sums to zero. This is a 5 parameter model, which, for simplicity we will reduce to a 3-parameter model by assuming symmetry between the nucleotides A and T and between the nucleotides C and G, that is, we set $\pi_A = \pi_T$ and $\pi_C = \pi_G$.

The stationary distribution of the matrix $P_{\mathbf{ij}}$ defined by Equation (3) for the HKY matrix with parameters $\alpha = 0.2$, $\beta = 0.5$, $\pi_A = \pi_T = 0.2$ and $\pi_C = \pi_G = 0.3$ for an effective population size $M = 30$ is illustrated in Figure 2. Equation (1) implies that the argument of the stationary distribution $\phi(i_A, i_C, i_G, i_T; \infty)$ lies on a simplex which, for a 4-letter alphabet, can be represented as a tetrahedron. In Figure 2 the distribution is represented by a small sphere at each set of integer coordinates, the volume of each sphere being proportional to the probability mass function at that coordinate.

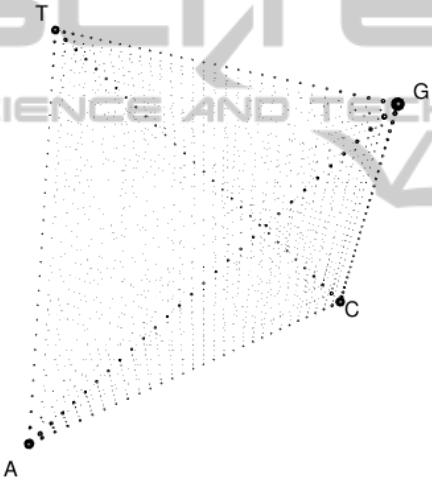

Figure 2: Stationary distribution of allele frequencies for the HKY model with parameters $\alpha = 0.2$, $\beta = 0.5$, $\pi_A = \pi_T = 0.2$ and $\pi_C = \pi_G = 0.3$. The corners labelled A, C, G and T correspond to the coordinates $\mathbf{i} = (1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$ and $(0,0,0,1)$ respectively, and the volume of the sphere at each coordinate point is proportional to the probability mass function.

The distribution is dominated by the corners of the tetrahedron, indicating that the majority of genomic sites are not SNPs. Edges of the tetrahedron correspond to 2-allele SNPs, the interiors of the four faces correspond to 3-allele SNPs and the interior volume of the tetrahedron corresponds to 4-allele SNPs. As is observed in real genomes, 3- and 4-allele SNPs are extremely rare. For illustrative purposes the parameter α has been chosen larger than what one might expect in nature by a couple of orders of magnitude to ensure the SNP probabilities are visible in the figure. The stationary distribution along the edges for each of the six possible 2-allele Figure 3 SNPs is plotted in Figure 3.

In the above example, the small effective population size $M = 30$ was chosen to enable a numerically tractable solution. For more realistic values of M the size of the matrix $P_{\mathbf{ij}}$ grows asymptotically as M^4. However, this does not present a problem as is should be feasible to develop a continuum Fokker-Planck equation analogous to Equation (5) with a tactible solution analogous to Equations (6) to (8) from which to calculate a log-likelihood.

6 DISCUSSION AND CONCLUSIONS

We have proposed an approach to estimating mutation rates from observed allele frequencies across a population at any set of independent genomic sites which are believed not to be susceptible to the effects of natural selection. Our numerical estimates based on the canonical textbook Wright-Fisher model of population genetics suggest that reasonable estimates of mutation rates can be obtained from as few as $\sim 10^4$ such sites.

The next step in this analysis is the technical problem of extending the continuum Fokker-Planck equation from the 2-letter genomic alphabet described in Section 3 to an analogous equation defined over the higher dimensional simplex relevant to the 4-letter genomic alphabet described in Section 5, and solving to find the steady state solution. This should be straightforward, at least for the Wright-Fisher model, and will enable maximum likelihood estimates of evolutionary rate matrices specified by parameterisations such as the HKY model. Going beyond Wright-Fisher to deal with species which are not diploid and monoecious will presumably not present insurmountable challenges provided the appropriate functions $a(\theta)$ and $b(\theta)$ analogous to those occurring in Equation (5) can be modelled.

An issue not addressed here at any level of rigour is that of context-dependence. As mentioned in the introduction, neutral mutation rates at a given site are known to be subject to the nucleotide content of flanking bases. We have assumed that the set of independent sites used to estimate rates are simply chosen to share the same context. However, this ignores the fact that the flanking bases may themselves mutate, and do so over timescales similar to the mutation rates we seek to estimate. We end on a note of caution that developing a non-local model which takes this into account may prove to be a formidable mathematical problem.

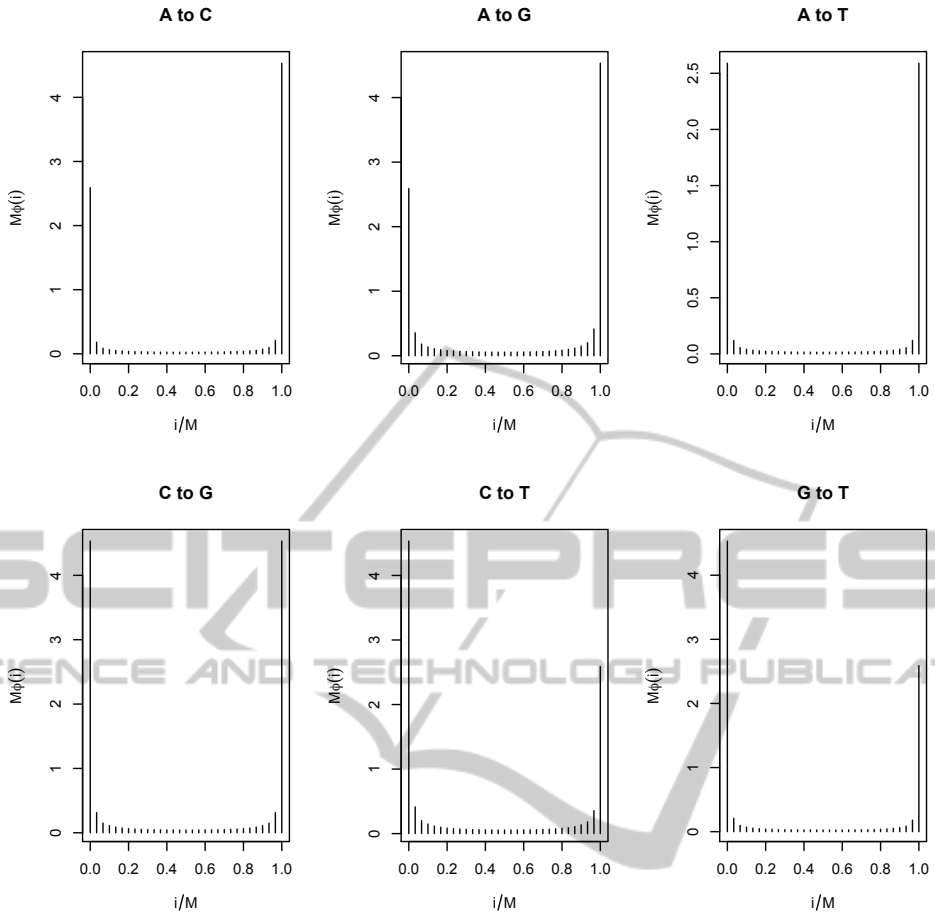

Figure 3: Distribution along each of the edges of the tetrahedron in Figure 2 illustrating the stationary distribution of allele frequencies in 2-allele SNPs within a population generated from the HKY model.

REFERENCES

Ewens, W. J. (2004). *Mathematical population genetics*, volume 27 of *Interdisciplinary Applied Mathematics*. Springer, New York, 2nd edition.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J Mol Evol*, 22(2):160–74.

Johnson, N. L., Kotz, S., and Kemp, A. W. (1992). *Univariate Discrete Distributions*. Wiley, New York, 2nd edition.

Lynch, M. (2009). Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics*, 182(1):295–301.

Messer, P. W. (2009). Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. *Genetics*, 182(4):1219–32.

Nevarez, P. A., DeBoever, C. M., Freeland, B. J., Quitt, M. A., and Bush, E. C. (2010). Context dependent substitution biases vary within the human genome. *BMC Bioinformatics*, 11:462.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16:97–159.

Zhao, Z. and Boerwinkle, E. (2002). Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res*, 12(11):1679–86.